Research on aircraft automation warning model based on gradient boosting tree algorithm

Xingmin Liu *, Bowen Su, Caiyi Zhang, Haixu Liu

School of Intelligence and Information Engineering, Shandong University of Traditional Chinese Medicine, Jinan, China

* Corresponding Author Email: 15689403550@163.com

Abstract. Flight safety is the basis for the survival and development of civil aviation transportation industry, and with the rapid development of China's civil aviation industry, the study of aircraft safety issues is becoming more and more important. Serious flight accidents can lead to great threat to life as well as economic loss, so this paper focuses on flight safety issues, and analyzes and investigates flight key parameters segment data, aircraft overruns, and flight parameters data. In this paper, based on the relevant data from QAR , t-test was performed on the processed data through data analysis and data cleaning to ensure the reliability of the data. And the principal component components of the key data were extracted using the random forest algorithm, and the 6-bit eigenvalues were selected as the key parameters, which had the greatest influence on the landing G value of the aircraft. Then, the data set was divided according to the ratio of 8:2, and the pilot qualification was used as the label value, and the machine learning classification prediction was performed by the gradient boosting tree GBDT algorithm to obtain the pilot qualification prediction model, and the highest accuracy reached 76.3% in the test set. This paper also compares the extreme gradient boosting XGBoost algorithm, with the highest current parameter reaching 73.4%. In this paper, the gradient boosting tree GBDT is chosen to construct a flight crew qualification assessment model.

Keywords: Flight safety; RF; GBDT; automated warning models.

1. Introduction

Currently, most of the flight technology assessment results are only simple words, lacking actual data support, so it is difficult to apply them in actual technology. Therefore, it is necessary to develop an accurate, reliable and practical assessment system evaluation, so as to reduce flight risks and provide an additional safety guarantee [1].

In landing safety analysis, landing G-value is usually an important indicator to describe the safety at the moment of landing. By modeling, analyzing and calculating the data when the aircraft is malfunctioning or failing, researchers assess the risk tendency, carry out targeted safety management, identify safety hazards and improve safety performance [2].

Based on this, this paper designs a predictive model based on aviation safety technology assessment through various machine learning algorithms [3]. In this paper, we first do reliability research on the quality of the given data after pre-processing a large amount of data, and on this basis, we process abnormal data to give quantitative results of flight safety factors and quantitative descriptions of flight operation data, and statistical analysis of overrun data, and build models based on pilot qualification assessment respectively. Finally, the above models are model integrated to establish an automated early warning model.

2. Model Preparation

2.1. Data pre-processing and quality analysis

QAR Quick Access Recorder, this data mainly records various flight parameters during the aircraft flight, including the aircraft's flight date specific time, altitude, descent rate, radio altitude, calculated

airspeed, ground speed, and air-ground electric gate, landing G value, attitude, stick volume, slope, disk volume, chute deviation and other data volumes related to flight to chicken.

After the big data statistics, the anomalies of the data are mainly generated by the missing and abnormal, in order to avoid the impact of distorted data on aircraft safety, so the quality of QAR data is very important for the establishment and solution of mathematical models. We first performed data pre-processing and quality analysis on QAR data, and the specific operation process is as follows.

In this paper, we first preprocessed the eight table data, screened out the definite class variables and quantitative variables, and performed reliability analysis on the data to deal with missing data and outliers, etc.

In this paper, after merging all the tables, there are 205,093 data in each field, and only the field "GEAR SELECT DOWN (landing gear)" has missing values, corresponding to each table, this paper will count the missing rate of this field, and the missing rate formula is as follows:

$$s = \frac{n}{N} \times 100\% \tag{1}$$

Where s denotes the missing rate, n denotes the number of missing data, and N denotes the total number of data.

Since data anomalies are more significant to the instability of the research system, losing a lot of useful information or making the final results unreliable, this paper directly eliminates the missing discrete data and supplements the continuous data according to the interpolation method.

For the abnormal data, this paper adopts the combination of box plot and 3σ principle to screen out the ones with higher deviation in statistics, using the method of upper and lower quartiles and 3σ principle, we first set the upper quartile as $\alpha_{0.75}$, and the lower quartile as $\alpha_{0.25}$, and derive the normal data sample interval as $[low,\ high]$ according to the combination of upper and lower quartile as well as 3σ value. The specific formula is as follows:

$$low = 3 - \sigma * r + [\alpha_{0.25} - 1.5 \times (\alpha_{0.75} - \alpha_{0.25})] \times (1 - r) \tag{2}$$

$$high = 3 - \sigma * r + [\alpha_{0.75} + 1.5 \times (\alpha_{0.75} - \alpha_{0.25})] \times (1 - r) \tag{3}$$

Where σ is the variance and r is the custom activation ratio, which is set to 0.3 in this paper. According to this, some box plots are made in this paper as figures 1-4:

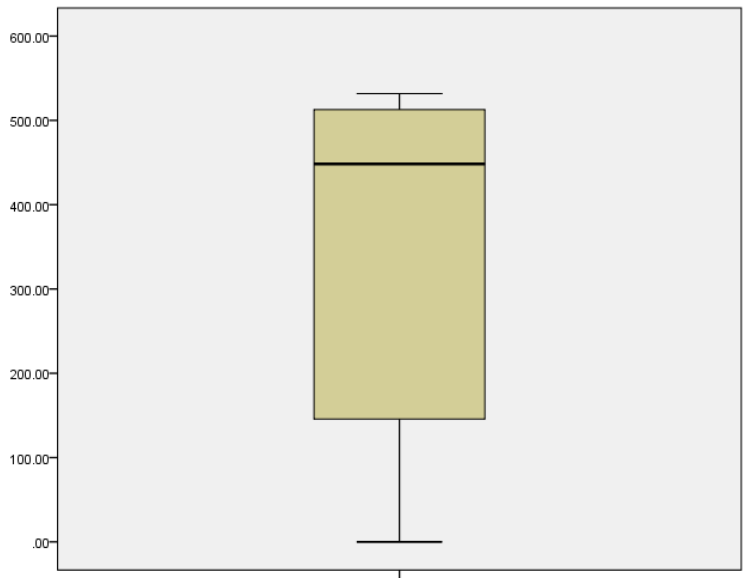


Fig 1. Ground speed box type diagram

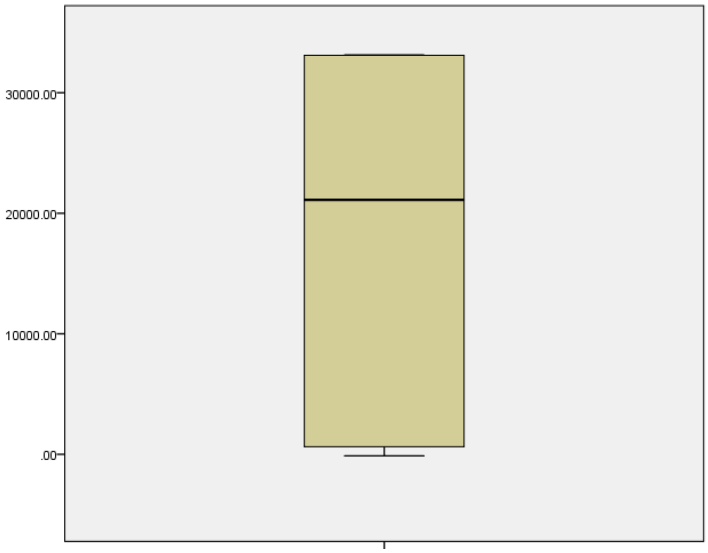


Fig 2. Altitude box plot

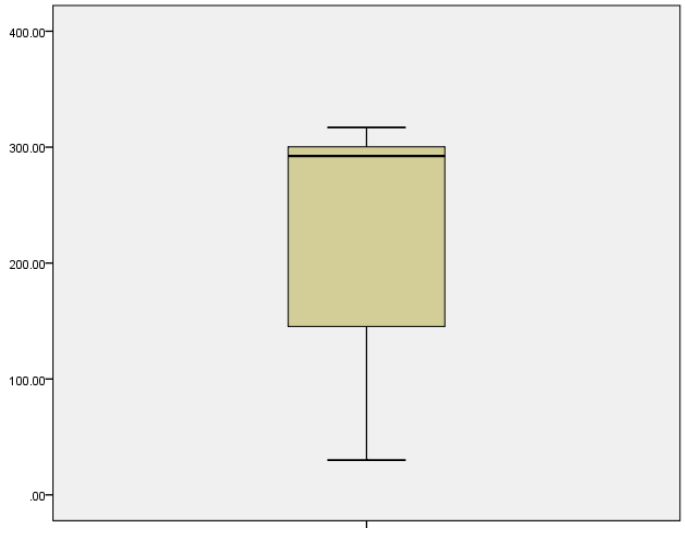


Fig 3. Calculate airspeed box type diagram

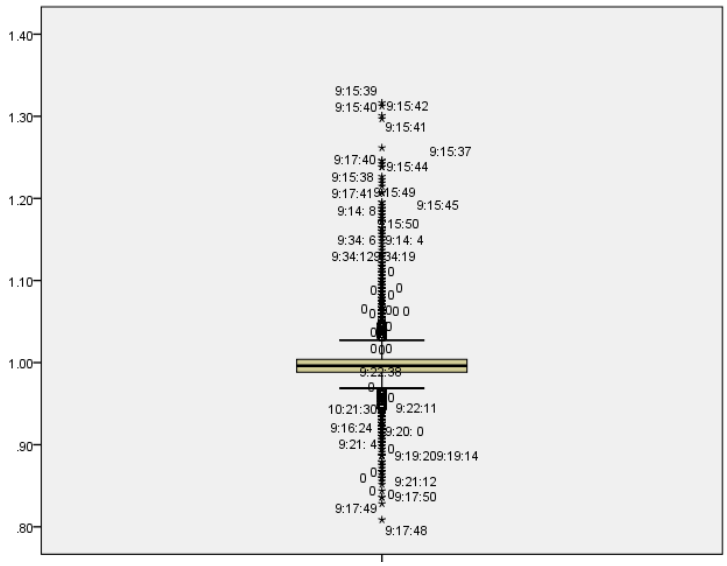


Fig 4. Landing G-value box plot

And in the box line plot, the part of the anomalous data greater than $high$ as well as less than low is replaced by the plural.

For the logical analysis of the data, this paper takes the logical relationship between altitude and aircraft descent rate as an example. The aircraft descent rate is the altitude that the aircraft descends in unit time, which is one of the important flight performance of the aircraft. The formula for the correlation between aircraft altitude and altitude is derived according to the definition as follows:

$$\mu = \frac{\Delta a}{\Delta t} \tag{4}$$

Where μ is the rate of decline, Δa is the elevation change per unit time, and Δt is the time interval.

2.2. Paired samples t-test

To ensure that the data are reliable and reasonable, the t-test is chosen to compare the data before and after processing. The differences between the two groups of data are compared to ensure that the data before and after the treatment still belong to the same group of data.

Table 1. T-test results

Variable Name	Sample size	Average value	Standard deviation	Skewness	Kurtosis	S-W test	K-S test
Before data processing	7461	1.002	0.029	1.623	9.123	-	-
After data processing	7461	0.998	0.022	0.856	15.78	-	-
Pairing	7461	0.004	0.037	0.508	6.611	0.888 (0.000***)	0.147 (0.000***)

Table 1 shows the results of descriptive statistics and normality tests of sample paired differences, including means and standard deviations, for testing the normality of the data.

● There are usually two tests for normal distribution, one is the Shapiro-Wilk test, which is applied to small sample data (sample size ≤ 5000), and the other is the Kolmogorov-Smirnov test, which is applied to large sample data (sample size > 5000).

● If it presents significance (P<0.05), it means that the original hypothesis is rejected (the data meets the normal distribution) and the data does not satisfy the normal distribution, and vice versa.

Usually it is difficult to meet the test in real research situations. If the absolute value of its sample kurtosis is less than 10 and the absolute value of skewness is less than 3, combined with the histogram of normal distribution, PP chart or QQ chart can be described as basically conforming to the normal distribution.

In this paper, the histogram of normal distribution is chosen for the description, as shown in Figure 5:

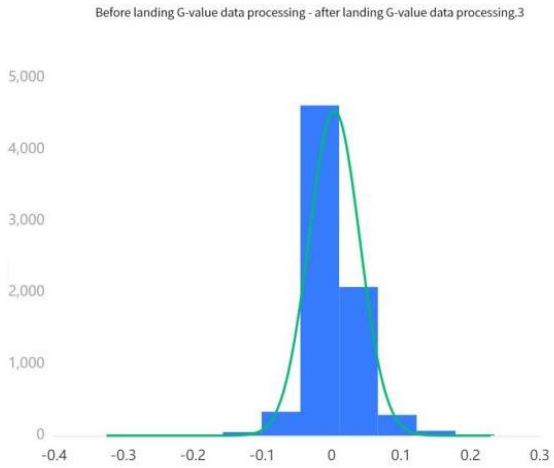


Fig 5. Histogram of normality test

Figure 5 shows the results of the normality test of the difference data of the quantitative variables before landing G-value data treatment and after landing G-value data treatment. If the normality plot basically presents a bell shape (high in the middle and low at both ends), it indicates that the data, although not absolutely normal, are basically acceptable as a normal distribution. According to the results of the paired-sample t-test in this paper, the variables landed G-value before paired landed G-value data treatment, the significance P-value < 0.5, the level does not present significance, and the original hypothesis cannot be rejected, so there is no significant difference between the paired landed G-value data treatment before paired landed G-value data treatment, and the data treatment results can be considered excellent.

2.3. Determining key parameters using random forests

Random Forest (RFF) consists of many Decision Trees, and there is no association between different decision trees. Decision Tree is an IF-THEN-ELSE algorithm that conforms to human intuitive thinking, and is a type of supervised learning algorithm. When performing a classification task, a new input sample is entered, and each decision tree in the forest is allowed to judge and classify that input sample, and each decision tree will get a classification result of its own, and which of the classification results of the decision tree has the most classifications, then the random forest will take this result as the final result [4].

The random forest algorithm has no obvious requirement on the input data dimensionality, no feature selection and feature value dimensionality reduction, and it is mostly used to judge the importance of different features and the interaction between different features. For the research problem of this paper, the random forest algorithm can well solve the problem of determining the key parameters. And when using random forest, it is not easy to overfitting phenomenon, and the accuracy can still be maintained in the face of feature loss.

In this paper, the maximum value of the interval from the landing G value of 0.1 seconds to the landing G value of 1 second is set as the target variable, and the variables with strong correlation are extracted by the Random Forest algorithm, and then the proportion of the influence of each variable on the target variable is output, which is arranged from the largest to the smallest in this paper.

By running the Random Forest algorithm on these parameters, the results of the importance of each parameter and its percentage.

Based on the information in the figure, this paper selects the parameters in the top 6 in importance as key parameters, i.e., pitch angle rate (PITCH ATT RATE), slope (ROLL ATT), pole volume mean (CAP CLM 1 POSN), calculated air speed (COMPUTED AIR SPD), ground speed (GROUNDSPEED), and descent rate (Inertial Vertical Speed) variables, which have the most influence on the landing G value of the aircraft.

3. An automated aircraft warning model based on gradient boosting tree algorithm

3.1. Prediction and classification model based on gradient boosting tree GBDT algorithm

Gradient Boosting Tree GBDT algorithm, GBDT is a decision tree based algorithm, which can fit complex linear and nonlinear relationships by constructing models through decision trees, and can be used for prediction and classification, which specifies weak learners as CART regression tree models. The core idea is to train the next weak learner by fitting the negative gradient of the previous model, so as to gradually reduce the bias. GBDT is mainly used to deal with structured and unstructured data, and can train an accurate model using limited data, and the resulting model has good generalization ability [5].

In brief, the gradient boosting tree GBDT algorithm is a combination of the gradient boosting algorithm and the CART regression tree algorithm. In the gradient boosting tree GBDT algorithm, the CART regression tree is used as a weak learner for data processing.

Using the gradient boosting tree, this paper makes the input-output process principle as follows:

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \tag{5}$$

Where 1 is used as the model input, 2 is used as the feature vector of the model, and each feature vector corresponds to a unique label value3.

After formatting the model input, the initialization of the weak learner is performed, and the weak learner initialization equation is as follows:

$$f_0(x) = \underset{c}{argmin} \sum_{i=1}^{N} L(y_i, c) \tag{6}$$

The iteration is executed M times, during which the following operations are performed:

$$r_{mi} = -[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f(x)=f_{m-1}(x)} \ (i = 1, 2, \dots, N) \tag{7}$$

The fit of the CART regression tree was performed using $(\boldsymbol{x}_i,\ r_{mi})$, resulting in m weak learners, where the regression tree corresponds to a leaf node region of $R_{mj}, j = (1, 2, \dots, J)$, where J is the number of leaf nodes of the mth regression tree. After completion, the best output value c_{mj} of R_{mj} is calculated as follows:

$$c_{mj} = \underset{c}{argmin} \sum_{x_i \in R_{mj}} L(y_i, f_{m-1}(x_i) + c) \tag{8}$$

After counting the best output values, the parameter values of the learner are updated using the parameters of c_{mj} to obtain $f_m(\boldsymbol{x})$, which is calculated as follows:

$$f_m(\boldsymbol{x}) = f_{m-1}(\boldsymbol{x}) + \sum_{j=1}^{J} c_{mj} I(\boldsymbol{x} \in R_{mj}) \tag{9}$$

The update of the corresponding parameter, $f_{m-1}(\boldsymbol{x})$ represents the model generated in front of it, which is also a step of the core idea of the reaction gradient boosting tree GBDT. This formula is used to continuously optimize the model. The final strong learner is obtained as shown below:

$$f(x) = f_M(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I(x \in R_{mj}) \tag{10}$$

Based on this, the gradient boosting tree GBDT algorithm is constructed in this paper to realize the classification fitting of flight personnel qualification. By continuously adjusting the model parameters to achieve better results, the relevant feature importance diagrams and heat maps are obtained after parameter adjustment as shown in Figures 6-8:

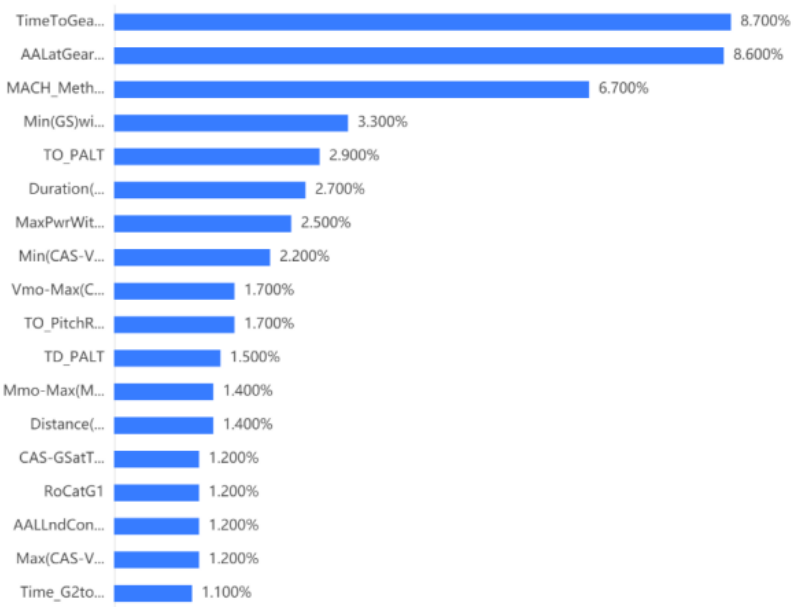


Fig 6. Correlation Eigenvalue Importance Chart

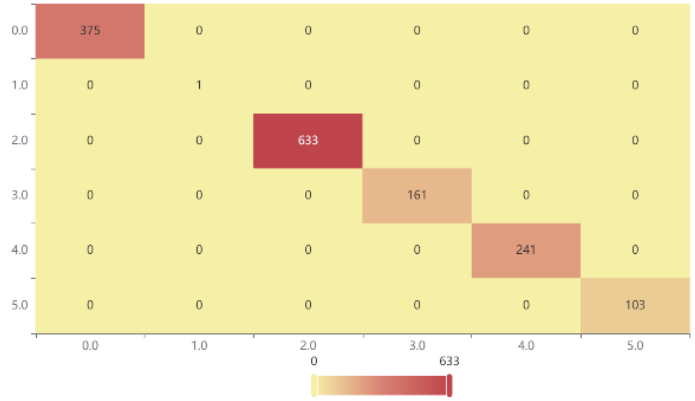


Fig 7. Confusion matrix heat map for training data

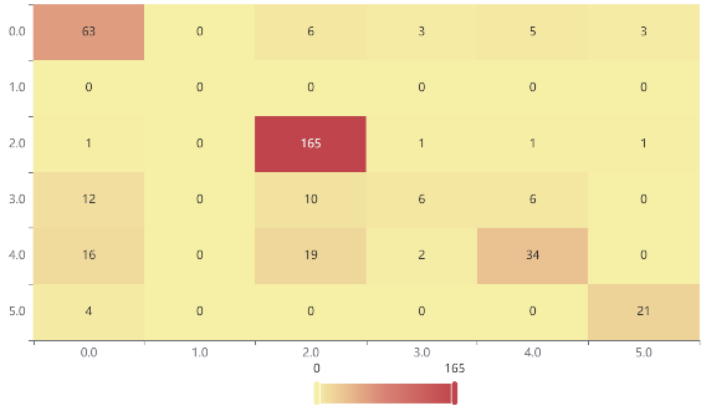


Fig 8. Test data confusion matrix heat map

3.2. Model evaluation and optimization

In this paper, the evaluation of the model is accomplished in various ways, including visualization of the model training process, evaluation of the model on multiple data sets, and comprehensive evaluation of the data during model training and testing.

In this paper, we visualize the model training process and obtain a graph of the variation of deviation with the number of learners, and we can see that the deviation value is decreasing and finally stabilizing, as shown in the figure 9:

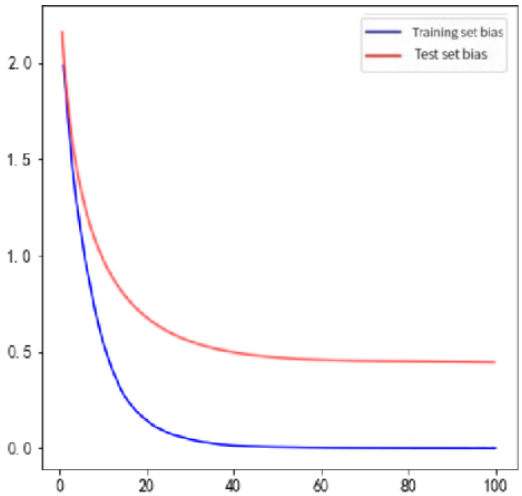


Fig 9. Variation of bias with the number of weak learners

In this paper, the model effect is evaluated by the accuracy of the model: the table 2 is obtained for the model evaluation data:

Table 2. Model Evaluation Data Sheet

	Accuracy	Recall Rate	Accuracy rate	F1
Training set	1	1	1	1
Cross-validation sets	0.686	0.686	0.661	0.655
Test set	0.763	0.763	0.745	0.738

In addition this paper shows the evaluation of the model results based on the data prediction tables during the training and testing phases (only some of the data are shown in this paper), and the test data prediction tables are as shown in tables 3 and 4:

Table 3. Data prediction table during training

Prediction Labels Y	Actual Labels Y	Predicted test result probability 0	Predicted test result probability 1	Predicted test result probability 2	Predicted test result probability 3	Predicted test result probability 4	Predicted test result probability 5
0	0	0.961129	4.12E-07	0.024309	0.010188	0.00431	6.35E-05
4	4	7.53E-05	4.51E-09	6.10E-06	8.53E-05	0.999833	4.78E-07
4	4	0.052161	4.89E-07	0.00067	0.004437	0.942678	5.47E-05
2	2	0.011162	2.70E-07	0.976567	0.00403	0.008211	2.98E-05
0	0	0.888171	1.57E-07	0.00021	0.110415	0.001187	1.66E-05
0	0	0.999937	3.38E-09	4.62E-06	2.11E-05	3.66E-05	3.57E-07
3	0	0.037079	2.09E-07	0.000266	0.960261	0.002371	2.22E-05
2	2	0.000498	1.94E-08	0.998993	0.000123	0.000208	0.000178
0	0	0.999228	2.62E-08	3.58E-05	0.000466	0.000259	1.09E-05
4	2	0.011578	1.22E-07	0.428799	0.004878	0.55473	1.35E-05
4	0	0.019619	2.70E-07	0.01166	0.005947	0.962737	3.70E-05
4	4	0.011318	2.56E-07	0.000871	0.007619	0.980165	2.71E-05

Table 4. Test Data Prediction Form

Forecast Y	Predicted outcome probability 0	Predicted outcome probability 2	Predicted outcome probability 3	Predicted outcome probability 4	Predicted outcome probability 5
0	0.883211	0.002756	0.075074	0.038652	0.000306
0	0.983239	0.000147	0.013801	0.002796	1.79E-05
2	0.046107	0.933056	0.004097	0.016639	0.000101
3	0.035192	0.000286	0.862411	0.101986	0.000125
2	0.005684	0.993483	0.000331	0.000497	5.39E-06
5	2.26E-06	3.31E-06	8.58E-07	1.89E-06	0.999992

In addition, there is room for optimization of the model: other parameters are selected for testing in this paper, which can make the accuracy of the model improve on the cross-validation set, but will correspondingly reduce the accuracy of the model on the test set, and these parameters are used to improve the conservative performance of the model in this paper, and the test results are shown in the table 5:

Table 5. Test Data Prediction Form

	Accuracy	Recall Rate	Accuracy rate	F1
Training set	1	1	1	1
Cross-validation sets	0.723	0.723	0.711	0.704
Test set	0.752	0.752	0.731	0.733

4. Summary

In this paper, a principal component analysis model was constructed by random forest, which has excellent computing speed and can quickly perform the analysis of key parameters and give a list of model importance. In addition, a pilot qualification evaluation model is established, and the pilot qualification is evaluated by the eigenvalue data with an accuracy of 76.3%, which can reach an even better accuracy through subsequent data filling and adjustment of model parameters.

The model proposed in this paper still has some room for improvement, for example, the proposed pilot qualification prediction model, which uses the gradient boosting tree GBDT algorithm to learn all the features after processing, can propose some features with small correlation for reinforcement learning in an appropriate amount to improve the accuracy of the model. In addition, this paper tries to use the extreme gradient boosting XGBoost algorithm for pilot qualification assessment and classification prediction, but does not achieve the desired results.

In this paper, we conducted an in-depth study on several parameters of aircraft flight and constructed an automated early warning model through data quantification, statistical analysis, random forest algorithm, and gradient boosting tree algorithm, which can effectively optimize the problems that occur during aircraft flight and enable aircraft pilots and towers to perform targeted operations for aircraft flight data. Through the migration application of the model. It can be effectively solved can be extended to include aerospace, marine navigation and other related fields with high requirements for safety, and has important reference value in the application process of practical problems.

References

[1] Liu JJ, Ye YH, Dong LY. Semantic mapping construction of aviation security information risk topics [J/OL]. Intelligence Engineering,2022(04):31-40 [2023-04-16].

[2] Liu, Mengna. Analysis of causal factors and risk prediction of aviation safety accident reports based on text mining[D]. Anhui University of Architecture,2019.

[3] Zhou Zhihua. Machine learning [M]. Beijing: Tsinghua University Press, 2016.01.

[4] Liu Yanli. A review of random forests [D]. Tianjin: Nankai University, 2008.

[5] Xu Yingzhuo,Guo Bo,Wang Liupeng. Research on game sales prediction model based on GBDT algorithm[J]. Intelligent Computers and Applications,2023,13(01):182-185.